November 2025

The FinLLM Journey: Building a Domain-Specific LLM for UK Financial Services

aveni

Table of contents

aveni

Introduction

The UK financial services sector is undergoing accelerated digital transformation, with generative AI emerging as a key enabler of efficiency, intelligence, and innovation. However, the deployment of general-purpose models in this tightly regulated environment has been limited by concerns around reliability, safety, and regulatory compliance.

The FinLLM project addresses these challenges by developing a domain-specific large language model tailored to the UK financial services industry. The model's capabilities and training strategy, from training data selection to evaluation, are designed to meet the sector's stringent governance requirements for data privacy, auditability, and responsible AI.

This paper describes:

- The motivations behind creating a domain-specific LLM for UK financial services.
- The process of data collection and curation for training such a model.
- The iterative approach to model development, from smaller prototypes to larger architectures.
- The validation framework used to ensure real-world utility and performance.
- The comprehensive safety, governance, and measures embedded throughout the model's lifecycle.

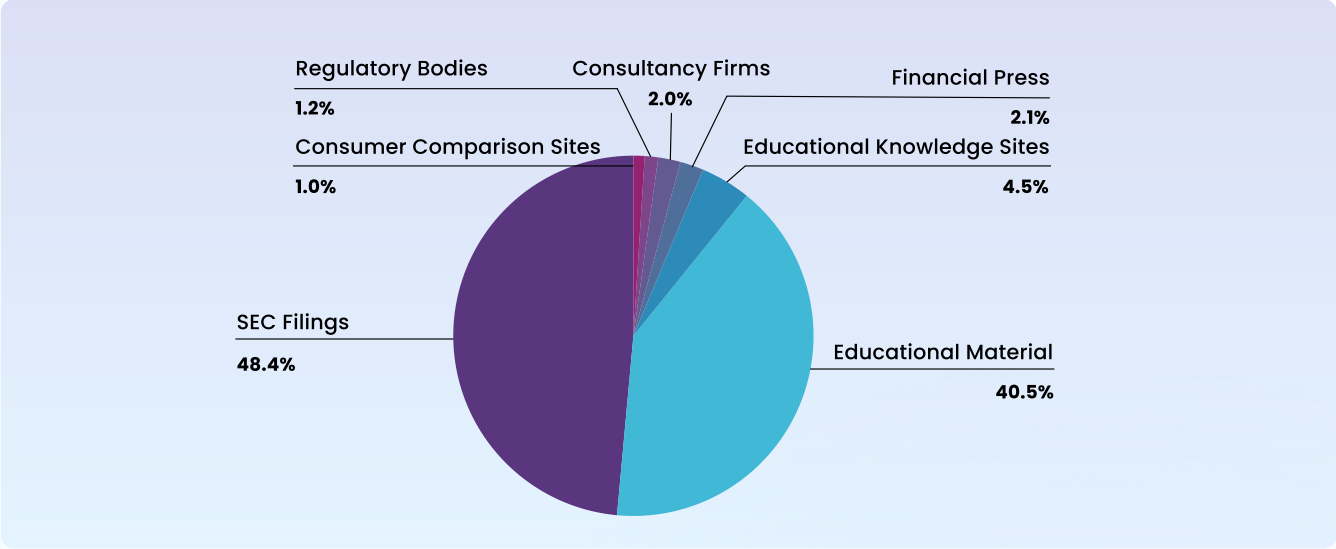
Chapter 1: The Foundation - Data

The performance of any specialised LLM is fundamentally tied to the quality of its training data. Recognising this, the FinLLM project dedicated significant effort to the creation and curation of **pre-training** and **post-training** data. **AveniVault** is a comprehensive **pre-training** dataset that has been ethically sourced and designed to serve as the project's training bedrock. All data is collected through responsible and transparent methods in full compliance with relevant regulations and guidelines. **AveniBlocks** is a collection of existing supervised fine-tuning datasets containing tasks covering finance, code, long-context and instruction following training datasets that deliver higher performance models for finance.

1.1 AveniVault Data for Continued Pre-training

The **AveniVault pre-training** dataset has evolved over the last 12 months in both size and features, expanding from an initial version (V1) of approximately 88 billion tokens to its current iteration (V2) of over 91 billion tokens. This growth was not merely in volume but also in strategic depth and diversity.

- **AveniVault-V1** established the foundation by combining web-scraped data from over 215,000 financial URLs with permissively licensed, large-scale corpora like HPLT and Fine Web, which were filtered for financial relevance.
- **AveniVault-V2** expanded this base with nearly 480,000 new URLs, adding approximately 3 billion new tokens. Crucially, this new data deliberately shifted the dataset's composition towards more structured and authoritative sources. **Educational material (40.5%) and regulatory documents (48.4%)**, such as SEC filings, FCA guidelines, and academic papers, now constitute the majority of new content, enriching the model's understanding of complex financial concepts and compliance frameworks.



Token distribution of newly added scraped websites in AveniVault-v2

1.2 A Multi-Stage Data Processing Pipeline for Pre-training Data

Data quality was prioritised through a multi-stage processing pipeline designed to clean, filter, and structure the raw data for optimal training.

Financial Classification: The largest proportion of financial data in the AveniVault comes from financial examples present in large data corpora. To surface this domain-specific data, **a financial classifier was deployed**. A pivotal enhancement was the development of our **Financial Classifier**. The classifier was trained using a more powerful teacher model and an efficient base architecture (ModernBert-base), which supports a large context length of 8,192 tokens. The result was a classifier that aligned very well with the teacher model and generalised effectively, achieving a Gold F1 score of 0.893. This directly translated into higher-quality financial data being fed into the models, boosting subsequent performance.

Cleaning and Filtering: The pipeline incorporates several key steps to ensure data integrity:

- **Content Extraction:** Parse HTML and PDF documents while preserving key metadata.
- **Toxicity Detection:** Label content for toxicity and its subtypes, allowing for the filtering of harmful text.
- **Deduplication:** Perform document-level deduplication, eliminating redundant content and improving training efficiency.
- **Language Identification:** Ensure the corpus remains focused on English-language content.

A Risk-Based Approach to Pseudonymisation: Data privacy is paramount in financial services. The project's approach to pseudonymisation matured significantly over time. The initial process used Microsoft Presidio for general PII detection. However, this was refined into a more sophisticated, **risk-based, three-tier system** linked directly to a newly developed financial data taxonomy. This was to avoid publicly-available and factual information being pseudonymised (e.g. senior business leaders, heads of state, etc) and to retain accurate information in the training data.

Data is categorised into levels, each receiving a different degree of pseudonymisation:

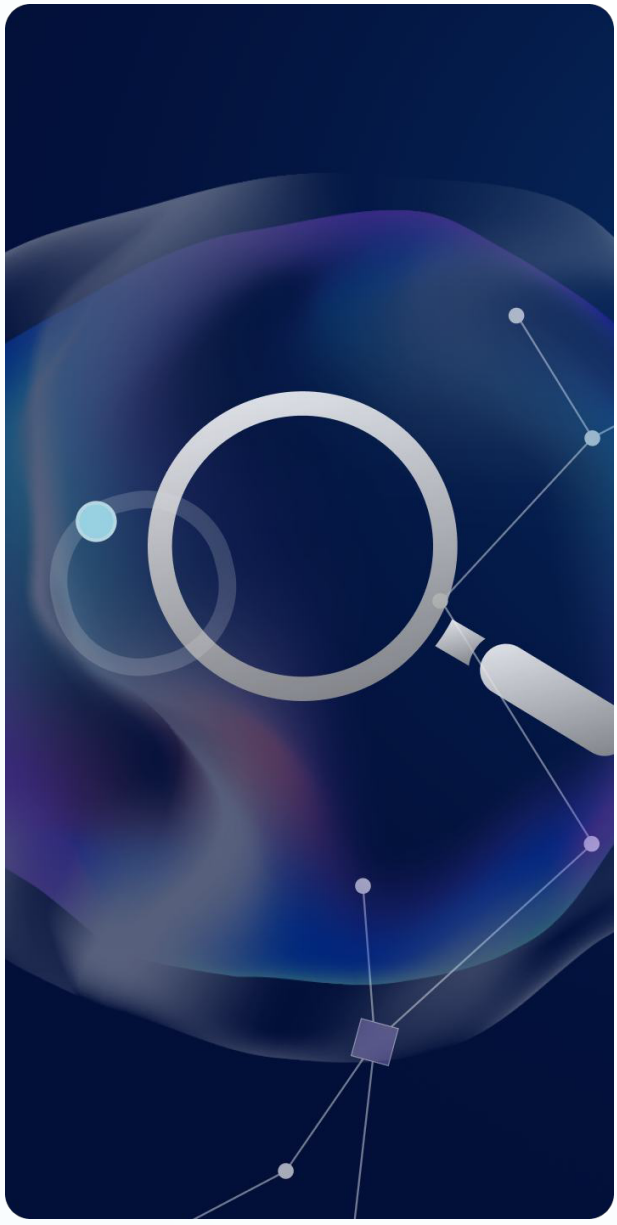
- **Level 0 (No pseudonymisation):** For public, non-sensitive information like academic content.
- **Level 1 (Financial data pseudonymisation):** For sources like financial news, where financial identifiers are redacted but personal names of public figures may be retained.
- **Level 2 (Full personal and financial data pseudonymisation):** The highest level, applied to sensitive sources like discussion forums, where all personal details are removed.

This nuanced approach allows the model to be trained on rich, factual information while rigorously protecting private data where necessary. Analysis shows a balanced distribution, with 48.4% of data at Level 0, 45.6% at Level 1, and only 6.0% requiring the strictest Level 2 pseudonymisation.

1.3 AveniBlocks Post-training Data

A high quality and balanced instruction dataset is essential to delivering models which perform well. We describe the progression of our instruction data from the initial version to the current v2.

- **AveniBlocks v1 Initial Mix (~50k Examples):** Our initial SFT datasets established a baseline mix, balancing finance-specific examples (from AveniBench training splits) with general instruction following math, tabular reasoning, and code data. A crucial finding from this stage was the importance of data diversity; even mixes containing significant non-finance data often *improved* performance on financial benchmarks compared to training on finance data alone.
- **AveniBlocks v1.5 Expanded Mix (~100k Examples):** To enhance performance on specific more challenging tasks (e.g. tabular and math) the dataset size was doubled for the next model iteration. This involved sourcing additional training examples for finance (e.g. converting more finance benchmarks into conversational format) and tabular reasoning. This allowed us to create a highly boosted finance-specific performance model (FinLLM-7B-Q-Tab).
- **AveniBlocks v2 Scaled & Enhanced Mix (~250k Examples):** For the FinLLM 24B model, the SFT dataset underwent its most significant expansion, growing 2.5 times larger. The mix was strategically enriched by adding an auxiliary MMLU training set (for broader knowledge and reasoning) and curating new, high-quality finance-related instructions from diverse sources filtered using our Financial Classifier. This scaled, quality-focused approach proved highly successful, directly contributing to the FinLLM 24B M model's superior performance, enabling it to surpass strong baselines across nearly all financial task categories.



Chapter 2: Building the Engine - A Journey of Model Iteration

The FinLLM project adopted a phased, iterative approach to model training, starting with smaller-scale models for rapid experimentation before scaling up to larger, more powerful architectures. This strategy allowed the team to test hypotheses efficiently and apply validated learnings at each successive stage.

Another fundamental aspect of our approach has been to deliver cost-effective training. We have strategically chosen performant open source models as our base models which we then efficiently adapt with Continual Pre-training (CPT) over AveniVault which imbues our models with foundational financial knowledge. Then we apply Supervised Fine-Tuning (SFT) with AveniBlocks instruction data, which transforms them from text completion models into helpful, instruction-following assistants capable of tackling real-world tasks.

Initially, we selected a family of models that demonstrated the strongest overall performance among all evaluated base options. We then incorporated an additional series of models developed within the European regulatory environment, making them particularly suitable for applications requiring strict compliance with regional AI governance standards. We also built on an open and fully transparent suite of models (those that clearly disclose their training data and methodologies) though these tend to underperform slightly compared to the other two base model types.

aveni

We started the project with the development of **FinLLM-1B-Q**, a smaller model (1.7 billion parameters) that served as a rapid prototype for testing data mixes and refining the training pipeline. In CPT we developed a two-phase training schedule and demonstrated the power of model merging to achieve robust performance.

Scaling up to 7 billion parameters, a 2-staged training approach yielded substantial improvements in the **FinLLM-7B-Q** models. Further specialised fine-tuning created **FinLLM-7B-Q-Instruct-Tab**, which achieved significant improvements in specialised financial tasks, particularly those involving tabular and mathematical reasoning. This breakthrough was largely due to the strategic use of synthetic data generation, which addressed the scarcity of high-quality training material in these areas.

The most recent work pushed the frontier with **FinLLM-24B-M**, notably outperforming all previous FinLLM versions and even the official base model across nearly every financial benchmark category. Additionally, recognising the need for efficient deployment, the team introduced FinLLM-14B-M, through a series of pruning-distillation-finetuning steps over the developed FinLLM-24B-M leading to a ~40% fewer parameters model than the 24B version while retaining highly competitive performance.

This results in a suite of models which are suitable depending on the performance, efficiency and compliance requirements.

2.1 AveniBench: A Comprehensive In-House Benchmark

To ensure consistent and relevant evaluation, we created AveniBench, an in-house collection of proprietary and permissively licensed datasets for benchmarking models on nine key NLP tasks critical for the finance industry. These tasks, identified through analysis of high-value industry use cases, include Text Classification, Long Context Modelling, Tabular Data reasoning, and Multi-turn Conversation. A public version of AveniBench containing eight datasets has also been released to the community, promoting transparency and standardised evaluation.

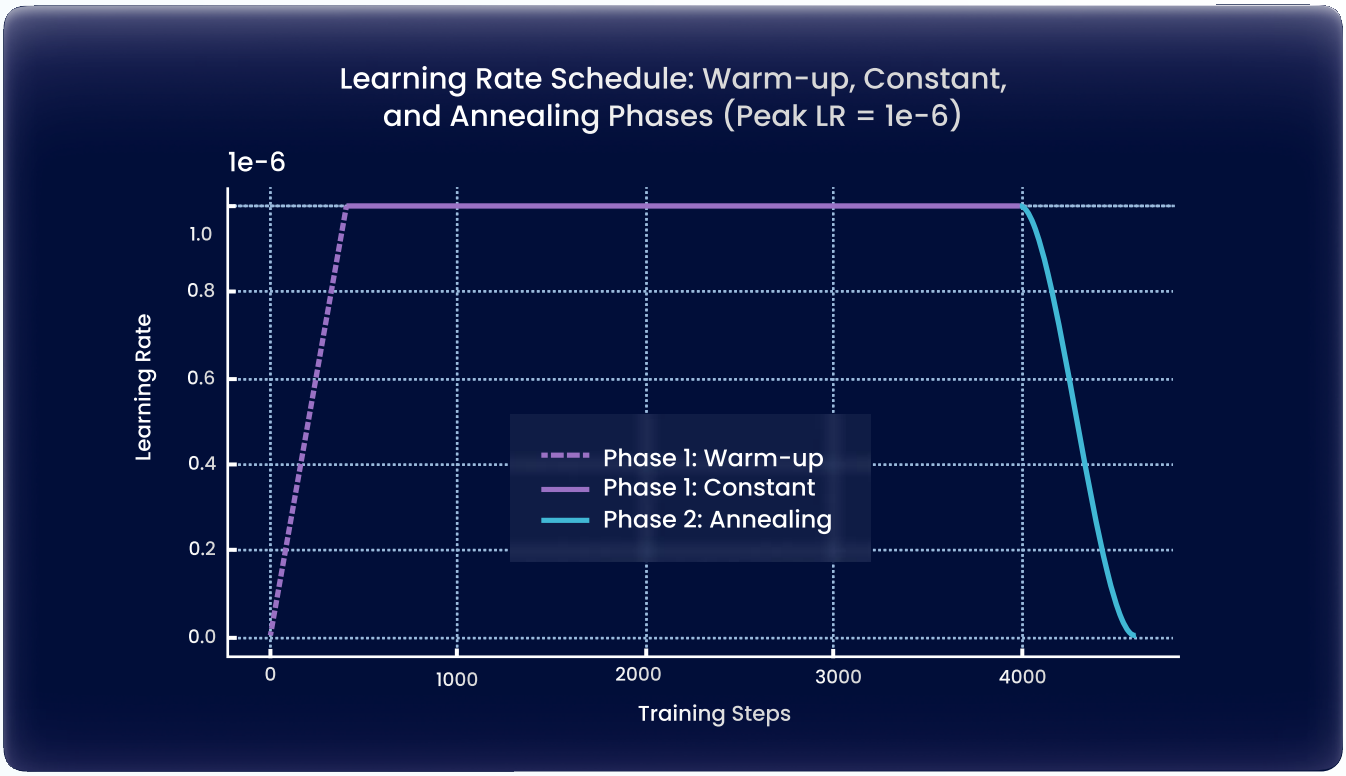
Our evaluation sets expand in three main categories: finance capabilities, general capabilities, and safety.

2.2 Rapid Prototyping with FinLLM-1B-Q

The initial experiments centred on a ~1B parameter model which served as an agile testbed for validating data mixes, tuning hyperparameters, and refining the training pipeline.

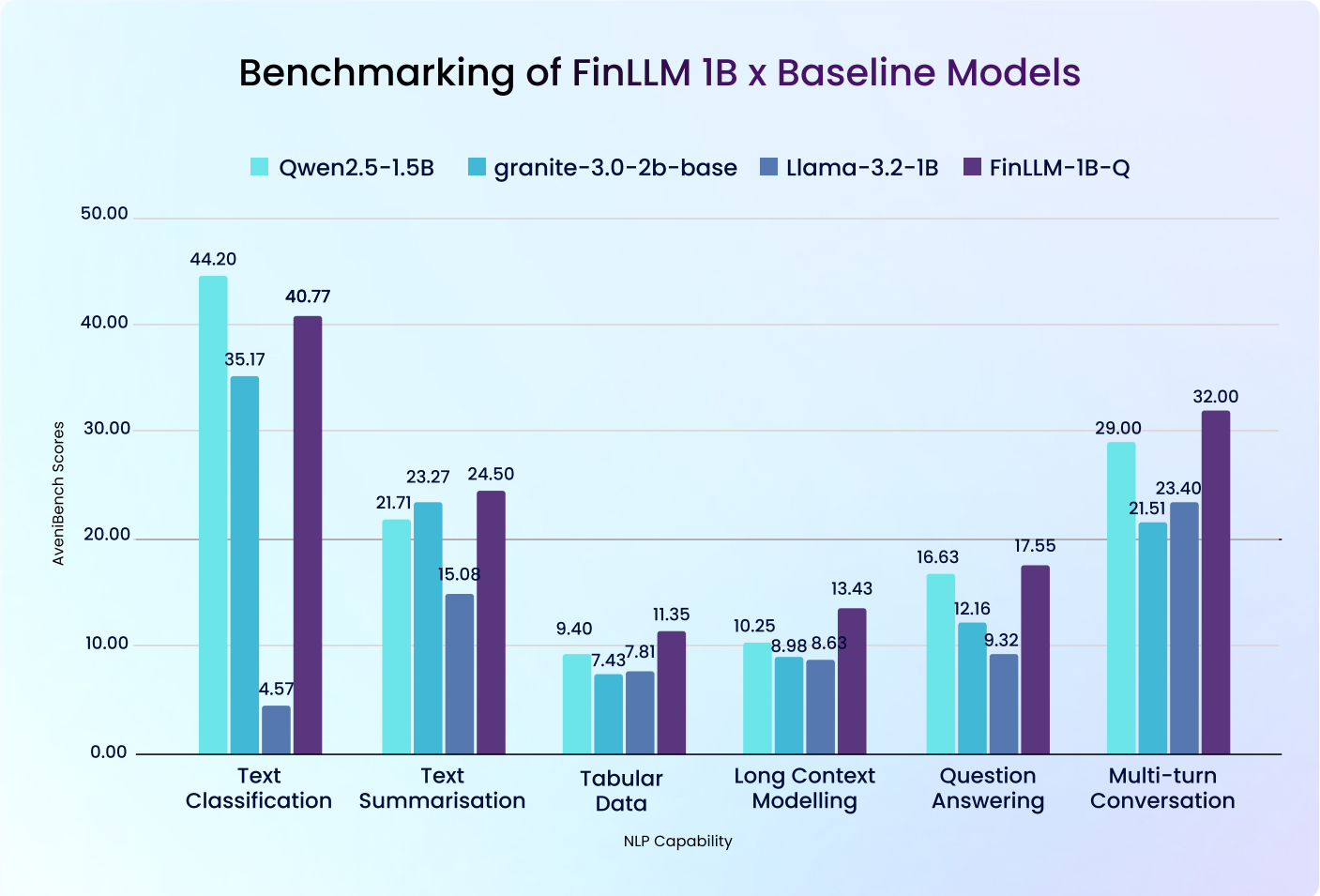
Two key learnings emerged from this phase:

1. **A Two-Phase Training Schedule:** A structured approach dividing Continual Pre-training (CPT) into two phases proved most effective for stable convergence. **Phase 1** involved a learning rate warm-up followed by a sustained plateau for steady learning. **Phase 2**, the annealing stage, used a decaying learning rate with a more targeted, high-quality data mix to precisely refine the model's financial knowledge.



The two different CPT training stages. In Phase 1 the learning rate is linearly increased to maximum, then kept constant for the majority of the training. In Phase 2 more targeted training data is used while decaying the learning rate (annealing).

2. Model Merging: The most robust 1B checkpoint was created not from a single training run, but by merging two different checkpoints using spherical interpolation (slerp). This technique effectively combined the complementary strengths of models trained with different learning rates and data mixes, resulting in the final, best-performing **FinLLM-1B-Q** model. The strongest pre-trained model was then fine-tuned to further boost its performance presented in section 1.3.



Performance of the best performing 1B base (non-instruction-tuned) model (FinLLM) against different targeted NLP capabilities. Similar sized baseline models are listed for comparison.

2.3 Scaling to 7B and the Impact of Synthetic Data

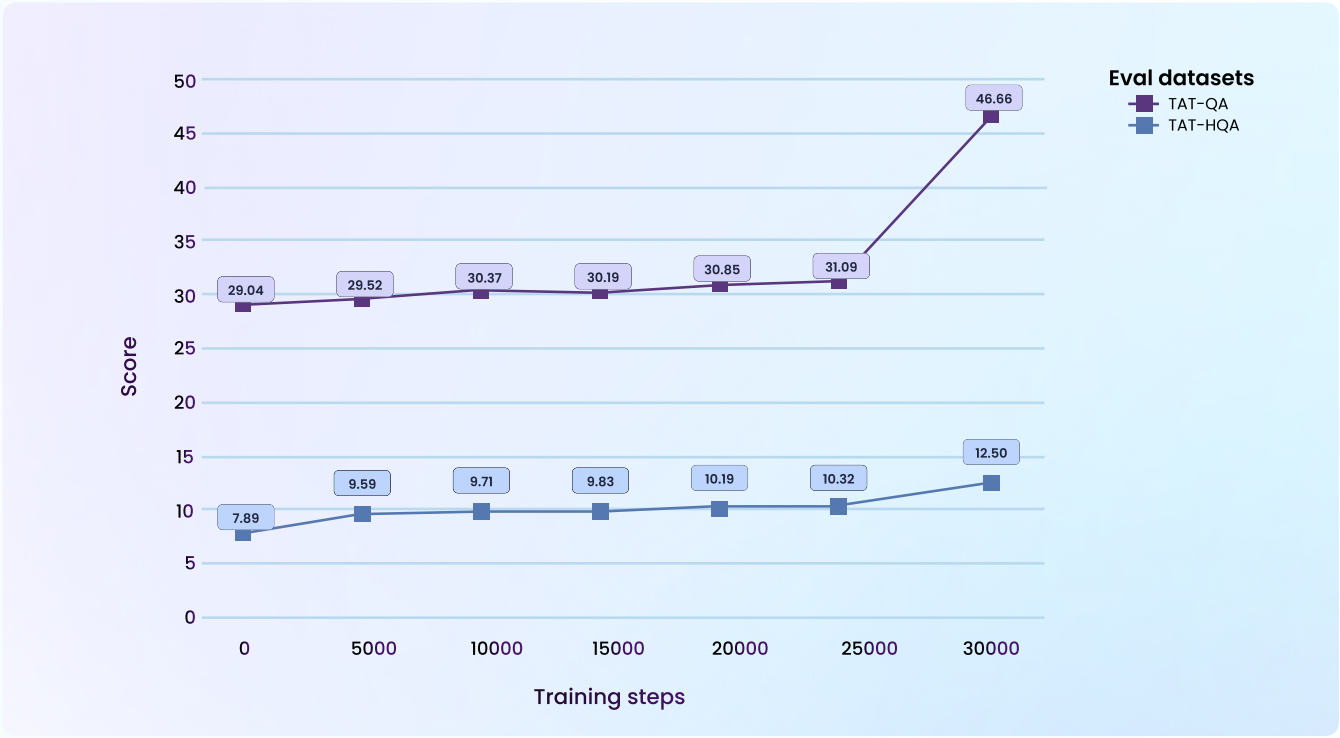
Armed with insights from the 1B experiments, the team scaled up to the 7B parameter class. The initial CPT run used 16B tokens and produced the FinLLM-1B-Q model. However, evaluation revealed a clear opportunity for improvement in specialised financial tasks, particularly those involving **tabular and mathematical reasoning.**

This challenge was met with a strategic pivot towards **synthetic data generation.** Acknowledging the scarcity of high-quality, large-scale tabular data, the team adopted a two-pronged approach to create over 2 billion tokens of new, targeted training material:

1. **Generating Educational Articles**: Inspired by projects like Cosmopedia, the team used LLMs to generate extensive, textbook-style financial articles from high-quality seeds (e.g. Investopedia, FCA guidelines). The prompts were designed to ensure the generated text included Markdown tables and calculations referencing them.

2. **Refining Existing Web Tables:** Starting with a finance-filtered subset of the WDC Webtables Corpus, an LLM was prompted to write analytical articles inspired by the original table and its surrounding text, explicitly requiring calculations and a summary.

During the initial CPT run, this data constituted only ~4% of the mix, leading to steady but modest gains. The breakthrough came during the subsequent annealing phase, where the synthetic tabular data was increased to **40% of the data mixture**. This targeted infusion of knowledge led to a dramatic and steep improvement in performance on tabular reasoning benchmarks.



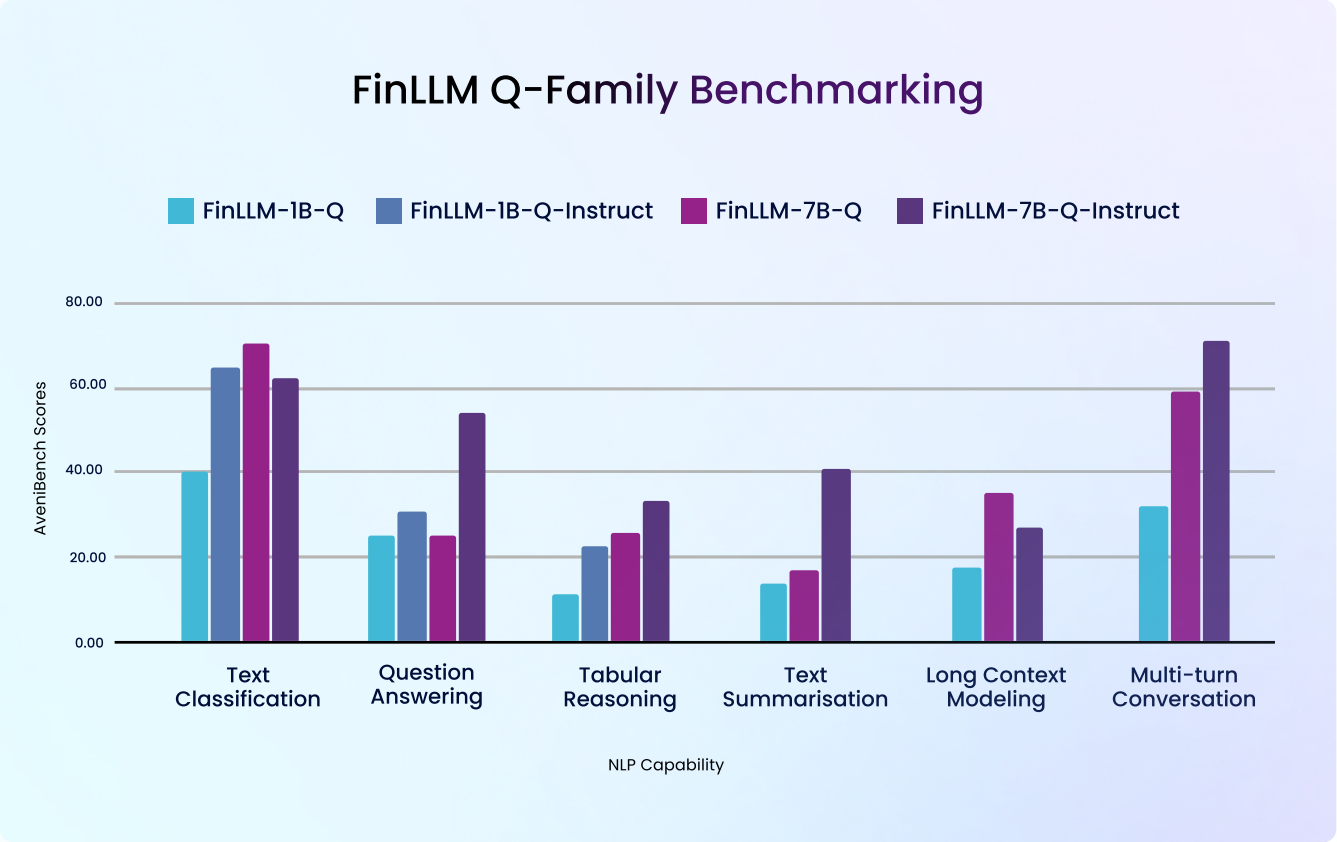
Intermediate checkpoint evaluation on tabular reasoning. The steep improvement of the last checkpoint is during the annealing phase with tabular data.

This process resulted in the **FinLLM-7B-Q-Tab** base model, which significantly outperformed its predecessor on math and tabular tasks.

This modular approach, particularly the ability to merge specific annealing runs, also provided a highly standardised and efficient process for fine-tuning models to excel in more pinpointed tasks.

Our SFT journey began with rapid prototyping on 1B parameter models. To accelerate iteration, we initially employed Low-Rank Adaptation (LoRA), a parameter-efficient technique. A key early learning was the necessity of unfreezing the final "LM head" layer alongside the LoRA adapters to ensure the model learned to generate proper end-of-turn signals, crucial for conversational applications. As we scaled to 7B models, we transitioned to **full-parameter fine-tuning** to maximise performance potential, while performing hyperparameter sweeping to identify the optimal learning rate.

As we scaled to 7B models, we transitioned to **full-parameter fine-tuning** to maximise performance potential, while performing hyperparameter sweeping to identify the optimal learning rate.



Performance of the FinLLM Q-Family base and instruct models against different targeted NLP capabilities. Note: Evaluations for text summarisation, long context modelling, and multi-turn conversation were not completed for the FinLLM-1B-Q-Instruct model.

2.4 Pushing the Frontier with 14B and 24B Models

The most recent work has focused on leveraging larger, more powerful base models and innovating on model efficiency.

FinLLM-24B-M: A new series of models was built upon an open source 24B base model. This FinLLM model has been generated by model merging of FinLLM 24B Single (a 24B baseline fine-tuned on the new and expanded SFT data mix as described in section 2.5) with a model that was first annealed before applying the same SFT. The results were unequivocal: the resulting **FinLLM-24B-M-Instruct** model not only surpassed all previous FinLLM versions but also outperformed the open source model across nearly every financial benchmark category. Apart from the fact that this provides us with a powerful ready-to-use model, it gives us with a strong baseline to train various smaller, more targeted models.

FinLLM-14B-M: Recognising the practical need for smaller, more efficient models for deployment, the team implemented an innovative pipeline based on the Minitron[1],[2] approach to create a high-performing 14B model from the 24B version. This multi-step process represents consists off:

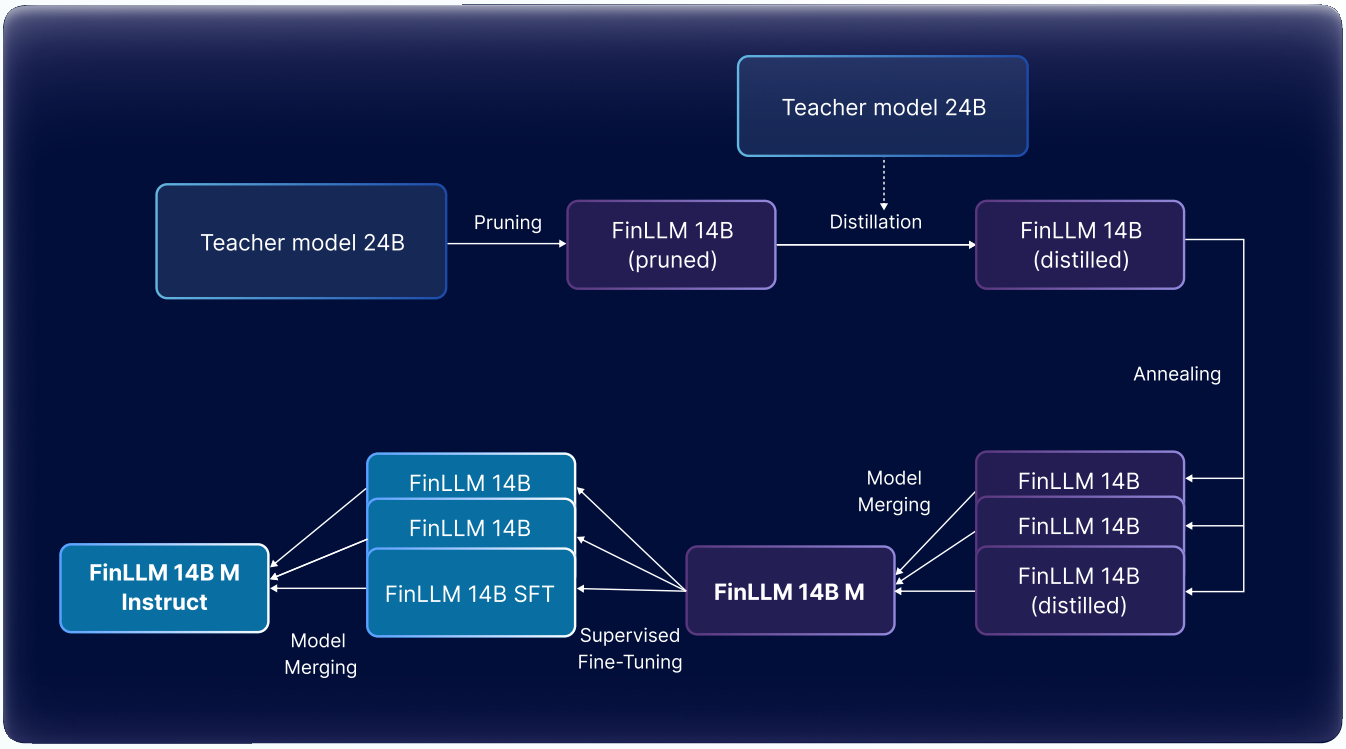
1. **Pruning**: The FinLLM-24B-M base model was reduced to 14B parameters through width pruning. This initial step drastically decreases performance as the model is losing knowledge due to limited parameters.

2. **Distillation**: Using the original 24B model as a "teacher," the 14B "student" model was trained on 5B general-domain tokens to regain the abilities lost during pruning, successfully closing a significant portion of the performance gap.

3. **Annealing**: The distilled 14B model was then infused with finance-specific knowledge through annealing, which further improved its performance on financial benchmarks beyond the distilled version.

4. **Model Merging & SFT:** Multiple annealing runs, which excelled on different tasks, were merged to create a superior base model. This was followed by a final SFT stage to produce the finished **FinLLM-14B-M-Instruct** model.

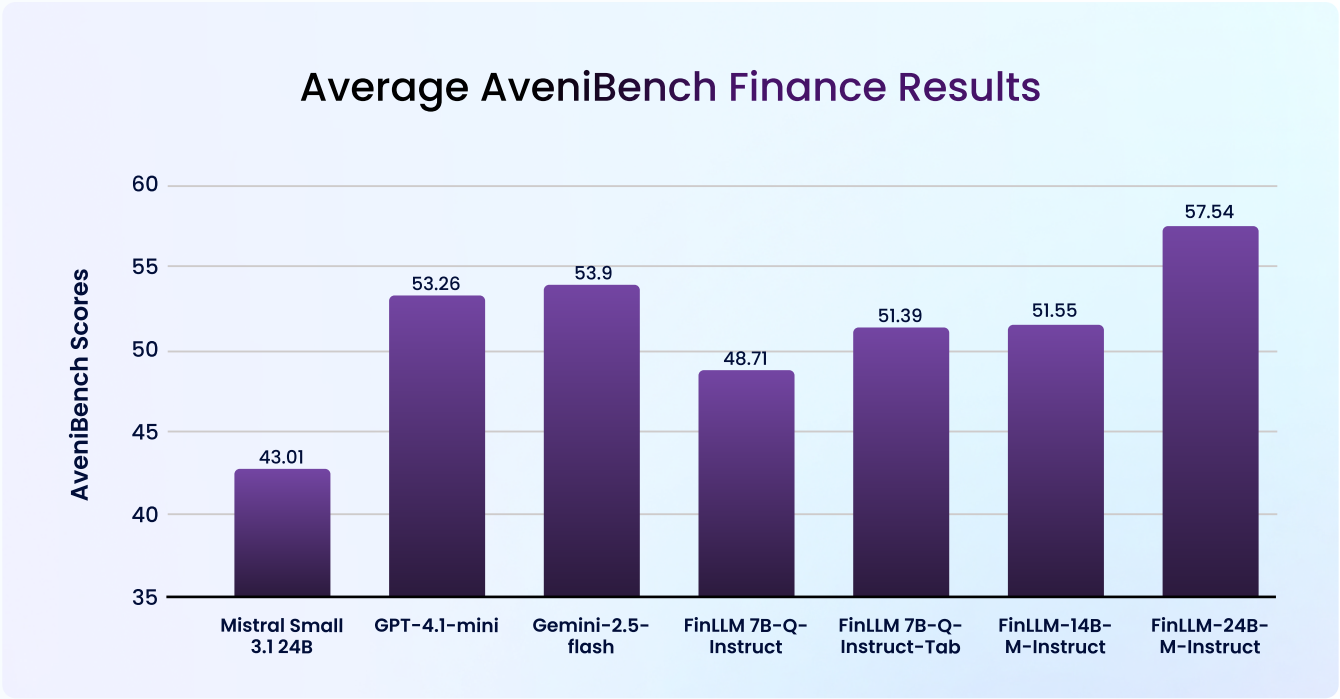
This pipeline successfully created a model with ~40% fewer parameters than the 24B version while retaining highly competitive performance, showcasing a practical path to efficient deployment.



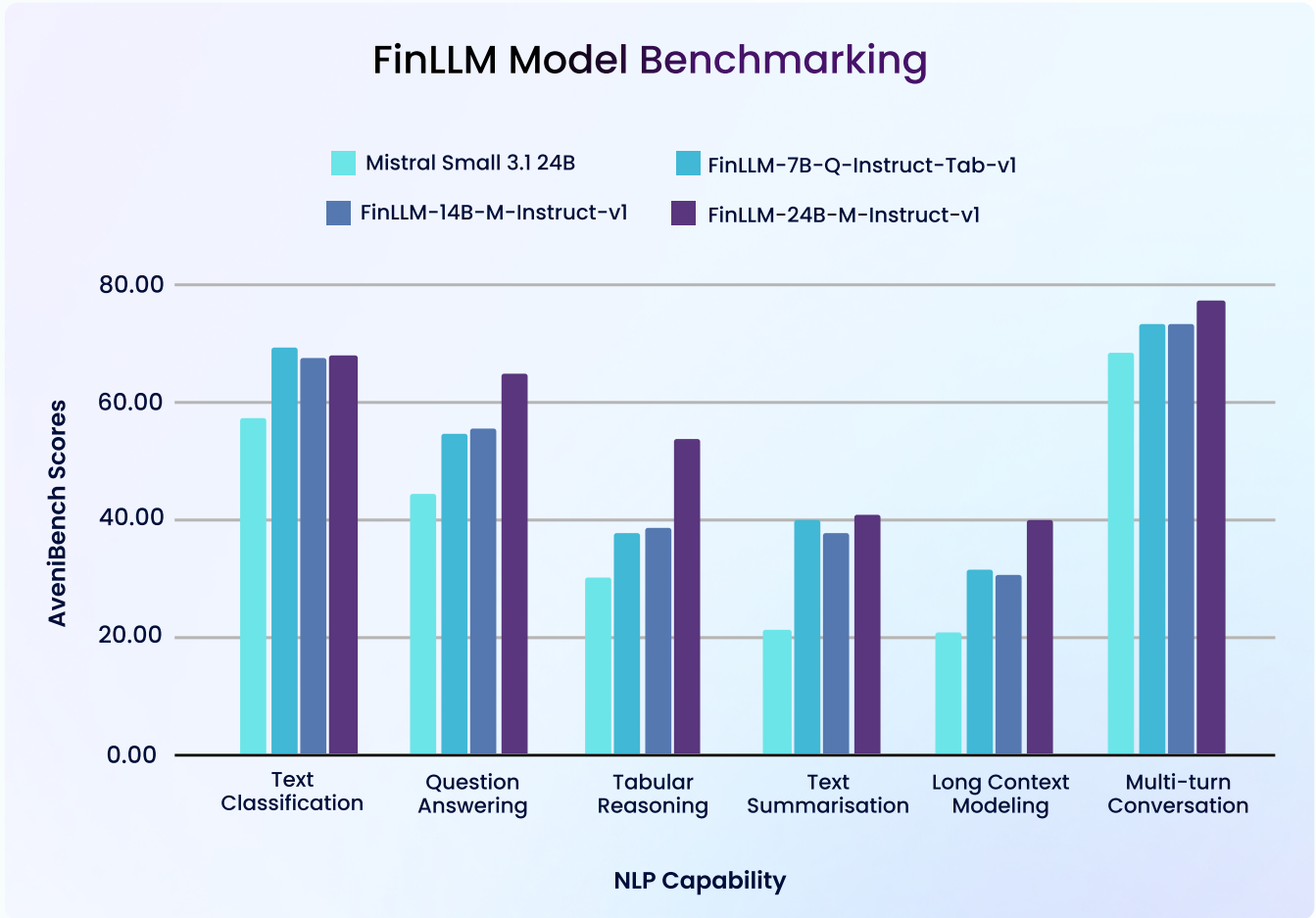
High-level overview of FinLLM-14B-M-Instruct pipeline. We start with the Mistral 24B model and through pruning, distillation, annealing, model merging and SFT steps build a final FinLLM model that has a size of 14B parameters.

[1] https://arxiv.org/abs/2407.14679
[2] https://arxiv.org/abs/2408.11796

aveni



In this graph we compare FinLLM to a number of very strong baselines, over the average scores in AveniBench. We show how FinLLM models of different sizes perform well, with FinLLM 24B demonstrating a clear competitive advantage.



The FinLLM models demonstrate a clear advantage in all 6 NLP capabilities in comparison to the similar-sized Mistral Small 3.1 24B.

aveni

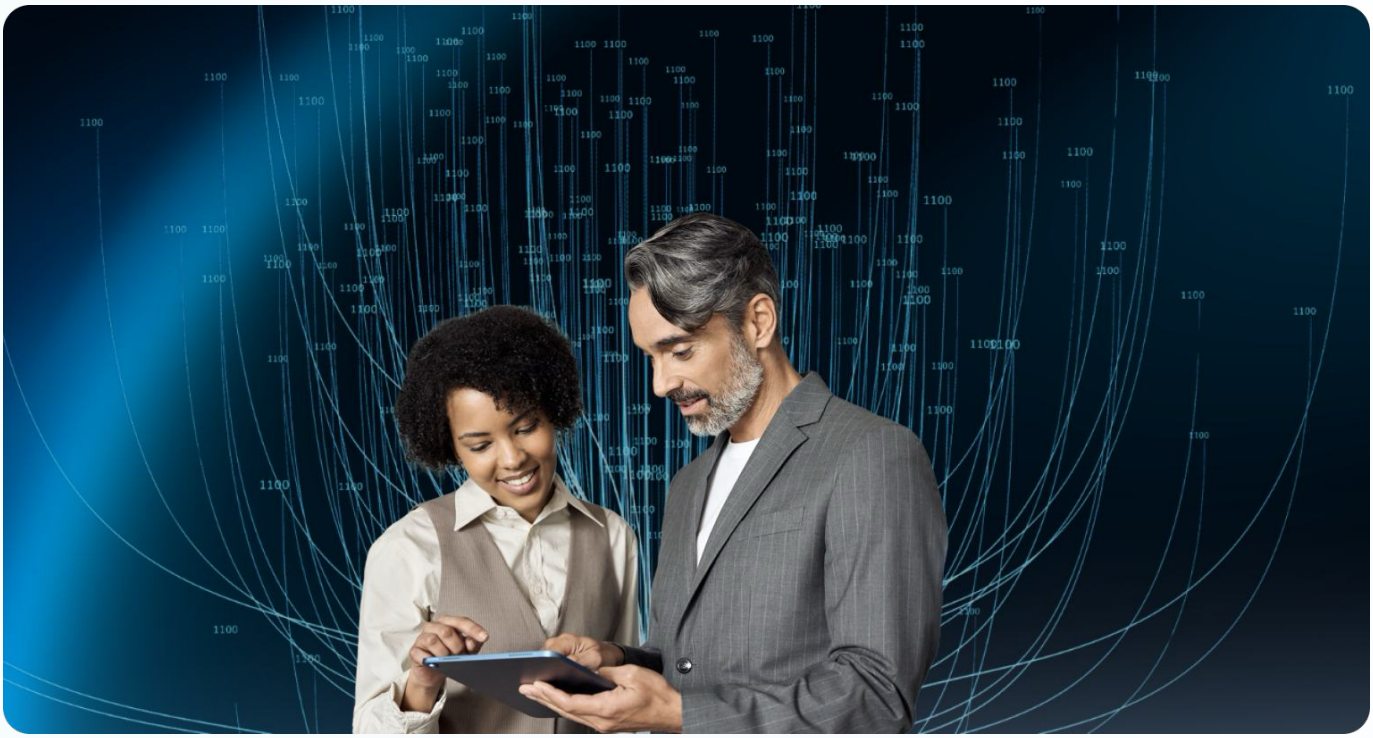
Chapter 3: Real-World Validation - Aveni Detect and Assist Use Cases

The ultimate measure of a model's worth lies in its real-world utility and its performance against comprehensive benchmarks, rather than solely on internal metrics. To this end, the FinLLM team implemented a rigorous evaluation framework, ensuring validation of its models throughout every development phase.

The most compelling validation of FinLLM's capabilities came from its application to Aveni's commercial platforms, demonstrating its practical utility in complex, real-world financial scenarios.

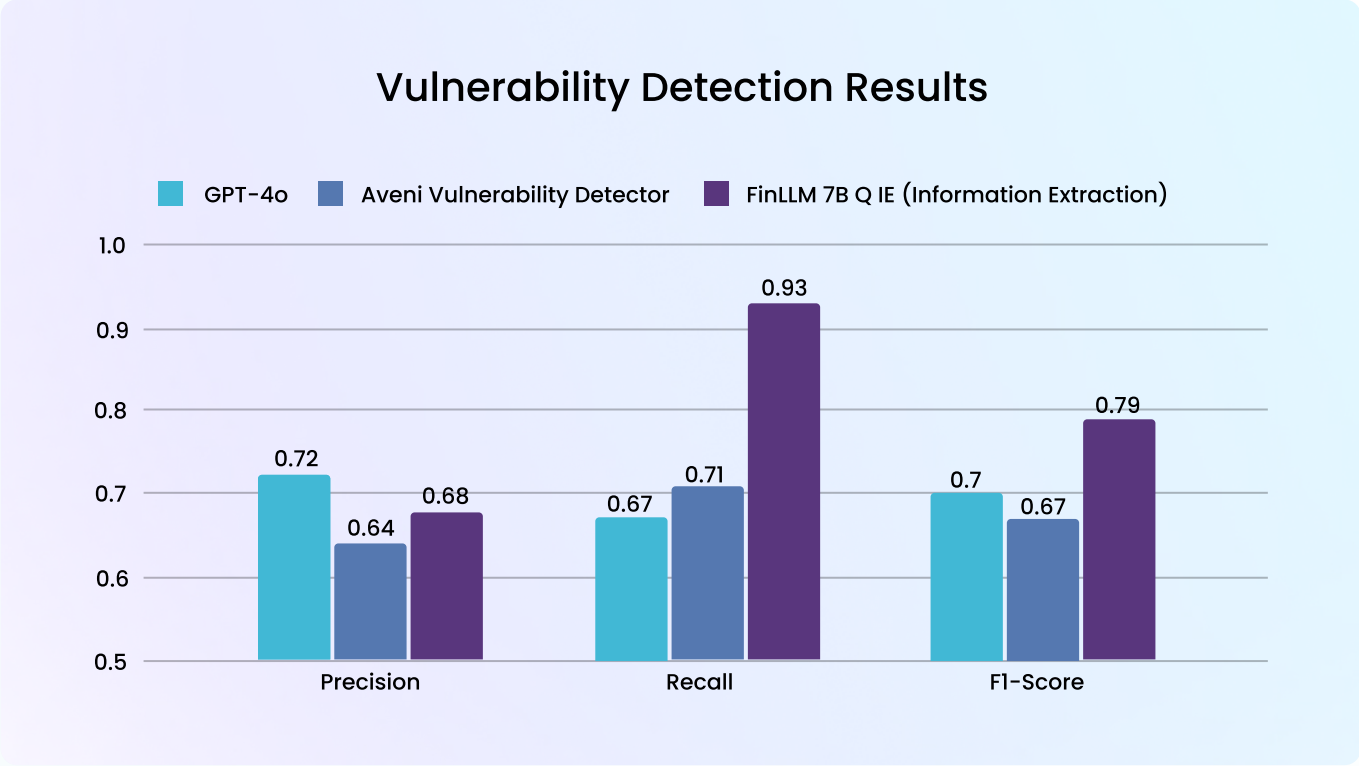
3.1 Aveni Detect: Simplifying Vulnerability Detection

Aveni Detect identifies risks such as customer vulnerability in lengthy adviser-client call transcripts. The existing production system uses a complex pipeline, splitting calls into 11-utterance windows and running them through a custom, RoBERTa-based vulnerability detector, where each window highlighted as a potential vulnerability is then sent to an API-based external LLM model for further classification.



The FinLLM workflow simplifies this process by using a single, fine-tuned FinLLM model to analyse the **entire call transcript** in one pass and sending the result to an external LLM, eliminating the need for splitting calls into windows of utterances and sending each one to an external LLMs. In an evaluation on real calls, **FinLLM-7B-Q-Instruct-IE** (FinLLM 7B fine-tuned on Information Extraction data. Details in an upcoming report) proved highly effective. In our evaluation we measure precision, recall, and f1-score at the call level – FinLLM classifies if this call contained a vulnerability, ignoring location FinLLM classifies which utterance(s) contains a vulnerability. **FinLLM achieved the highest F1 score** for identifying vulnerabilities, outperforming larger commercial models like GPT-4.1 Mini and GPT-5. It also demonstrated the highest **recall** across the evaluated models, a very important metric in detecting vulnerabilities, given that in any such use-case

This demonstrated that a smaller, use-case-specific model could deliver superior performance and efficiency, reducing both cost and complexity.



The results on the HITL[3] dataset, shown in the table above, shows that FinLLM 7B Q Instruct IE is competitive with OpenAI's GPT models (which have considerably more parameters), and significantly outperforms the similar-sized Qwen2.5 7B Instruct, Aveni's RoBERTa based vulnerability detector, and FinLLM 7B Tab Q. In terms of vulnerability classification, FinLLM 7B Instruct IE Q obtains the highest f1-score out of all models.

[3] Human-In-The-Loop dataset: Real customer calls that were annotated for vulnerability by users of the Detect Platform and have not been pseudonymised. These reflect the actual production calls with user labels the most.

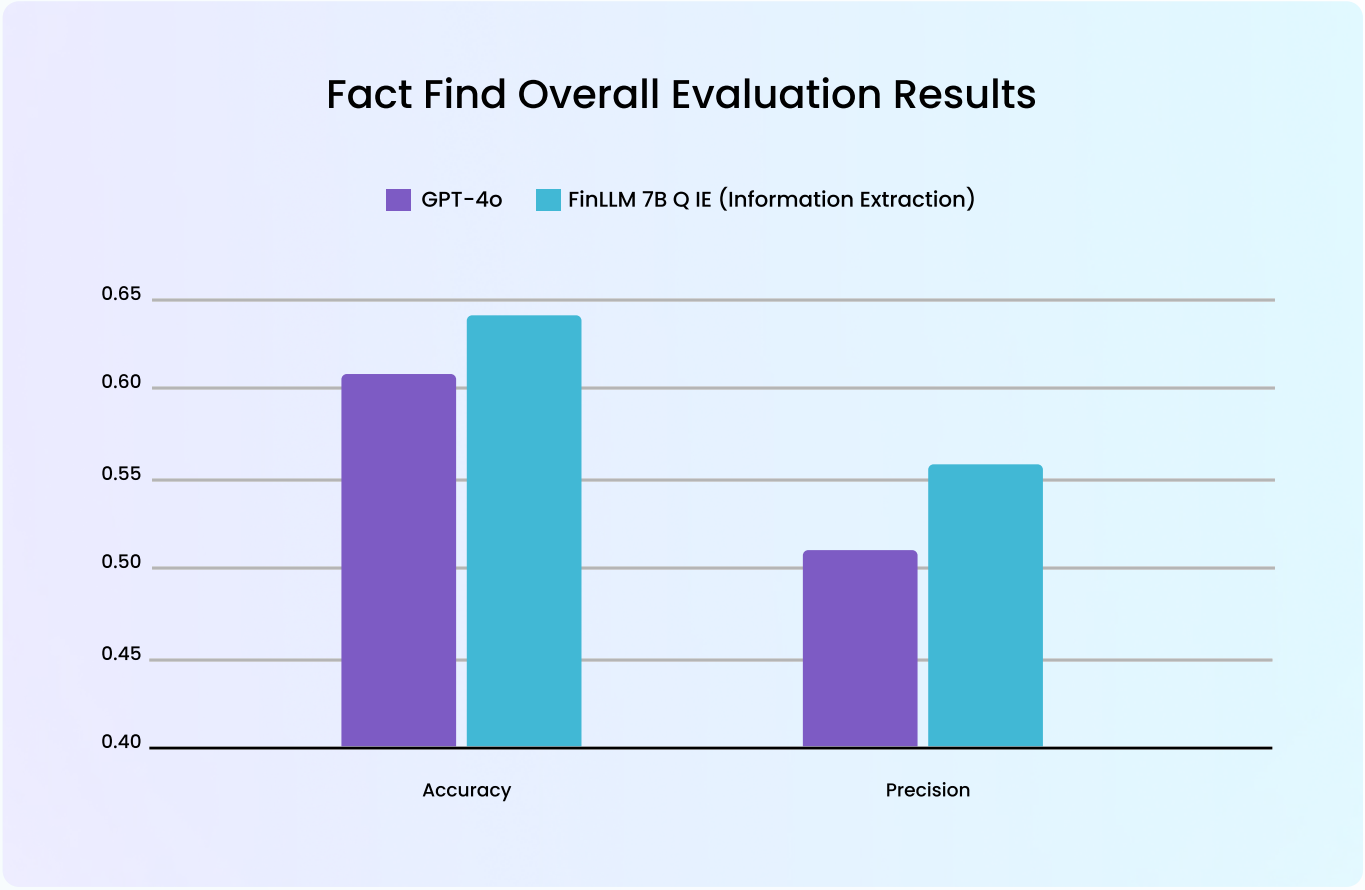
aveni

3.2 Aveni Assist: Enhancing Fact Extraction

Aveni Assist automates the extraction of structured financial facts (e.g. income, assets, expenditures) from call transcripts to populate a client "FactFind"[4].

The project aimed to replace parts of the pipeline with a fine-tuned FinLLM model. Due to the nuanced nature of the task, where outputs can be semantically correct but lexically different (e.g. "annually" vs. "yearly"), an evaluation method using an LLM-as-a-judge was employed to assess correctness. The results showed that each of the FinLLM models fine-tuned on information extraction and synthetic data outperformed GPT-4o in terms of accuracy (the primary metric), highlighting their suitability for use in production, thanks to the improved SFT data mix that includes synthetic use-case data, finance-specific instruction following, information extraction, and tabular QA. This proved that a targeted FinLLM could increase accuracy and reduce reliance on a more expensive, general-purpose API for a critical task.

These coarse-grained results, whilst useful for an overview, mask performance at the individual category and field levels.



[4] The output of the FactFind process is a table that contains factual information extracted from summaries of transcribed calls between an adviser and their client(s). For each category (income/asset/expenditure/etc.) the following information is extracted: name(s) of the client(s), the type of item (e.g. “salary” in the case of income), the value of the item including the currency, the frequency of the item (e.g. “monthly”, “annually” etc.), and any additional notes that are relevant. This information may then be manually corrected by the adviser before (optionally) being pushed to a CRM.

Chapter 4: Ensuring Trust - Safety, Governance and Compliance

For an LLM to be deployable in financial services, performance must be matched by an unwavering commitment to safety and ethics. The FinLLM project has embedded a multi-layered safety and governance framework throughout the entire model lifecycle.

4.1 A Multi-Layered Mitigation Framework

The project identified **eight primary risk categories**: Toxicity, Bias, IP Infringement, Privacy, Misinformation, Misalignment, Hallucination, and Sustainability. Mitigation strategies are applied at every stage:

- **Governance:** Maintaining transparent records of data sources, licences, and adhering to policies like the Data Protection Impact Assessment (DPIA) and aligning with the EU AI Act.
- **Pre-training:** Using model-based detection tools to identify toxic and biased content from CPT data and implementing the risk-based pseudonymisation pipeline to protect personal data.
- **Fine-tuning:** Incorporating specific datasets during SFT to improve alignment and reduce bias.
- **Evaluation:** Utilising a comprehensive suite of safety benchmarks (e.g., BBQ, BOLD, ToxiGen, TruthfulQA) and red-teaming to assess performance against each risk category.
- **Deployment:** Implementing a robust guardrail system to monitor model inputs and outputs in real-time.

4.2 Guardrails and Safety Performance

A context-sensitive, three-stage guardrail framework is being developed to balance safety with performance:

1. **Pre-call guardrails** check user inputs for the most severe risks, such as jailbreaking attempts or toxic prompts, and block them before they reach the model.
2. **During-call guardrails** provide a second threshold for inputs that pass the first check, stopping the model from generating a response to prompts that are borderline but still risky (e.g., potential misinformation).
3. **Post-call guardrails** assess the model's generated output for risks like hallucination, bias, or privacy leaks before it is sent to the user.

The project plans to use the NVIDIA NeMO Guardrail Framework and models such as LlamaGuard 8B for the implementation of this system.

Evaluations have shown that safety mitigations are effective. Using a simple system prompt during evaluation increased the average safety performance of FinLLM-7B-Q-Instruct from 62.33 to 67.69, outperforming other similar-sized models. While fine-tuning with specific safety training datasets (FinLLM 7B Instruct Tab Q v1.0.1) further increased the safety performance to 69.37. Specifically, we address toxicity and bias, misalignment, and jailbreaking.

These two methods of safety mitigation (system prompt and SFT) not only increased the average performance on safety evaluation benchmarks, but also increased both finance and general benchmark performance compared to FinLLM-7B-Q-Instruct.

Following the success of our 7B fine-tuning, in future we plan to apply the same approach to our 7B (Tab), 14B, and 24B models to improve their performance in the safety benchmarking.

FinLLM Safety Benchmarking by Safety Categories[5]

	FinLLM 7B Q Instruct	FinLLM 7B Q Instruct (Prompt)	**FinLLM 7B Q Instruct Tab v1.0.1**[6]	FinLLM 7B Q Instruct Tab	FinLLM 24B M Instruct	FinLLM 14B M Instruct	Mistral Small 3.1 24B
Toxicity	82.87	92.45	91.94	75.63	82.58	73.44	93.06
Bias	75.35	77.20	76.30	74.53	75.59	75.75	82.95
Hallucination	70.30	69.01	73.76	71.66	67.30	56.89	72.43
Misinformation	35.75	43.81	46.89	31.33	42.87	38.95	48.92
Misalignment	60.58	63.72	62.90	55.44	64.36	62.78	63.45
IP Infringement	77.00	96.00	99.00	64.00	66.00	74.00	86.00
Privacy	63.11	94.63	78.00	50.60	54.30	52.20	62.00
Safety Average	**62.33**	**67.69**	**69.37**	**57.84**	**64.14**	**60.59**	**68.10**

However, the project also transparently acknowledges its challenges, which remains a key area for future work through alignment techniques like Direct Preference Optimisation (DPO) and Reinforcement Learning from Human Feedback (RLHF).

Finally, in line with its commitment to responsible AI, the project actively tracks the **carbon footprint** of its training runs, reporting metrics such as total power consumption (130.93 tCO2 for all 7B continued pretraining runs) to ensure awareness of its environmental impact. As we continue our development of the FinLLM model suite we plan to expand our environmental impact reporting to provide carbon emissions of both training and inference.

[5] *A breakdown of the datasets within each safety category can be found in the FinLLM Safety White Paper: Part 2*
[6] Fine-tuned for Safety

Conclusion and Future Outlook

Over the past 12 months, the FinLLM project has progressed from concept to capability, establishing the foundations for a sovereign, domain-specific language model built for the UK financial sector. The work has proven that a smaller, specialised model trained on ethically sourced, regulatory-aligned data can outperform much larger general models in targeted financial tasks. Through an agile and iterative approach to model development, synthetic data generation, and rigorous evaluation, FinLLM has become a credible alternative to proprietary models - one that is transparent, explainable, and grounded in the realities of regulated AI deployment and an uncompromising focus on real-world validation.

Key achievements include:

- The creation and refinement of **AveniVault**, a 91B+ token dataset rich in UK-centric educational and regulatory financial content.
- The creation of **AveniBlocks**, a comprehensive collection of supervised fine-tuning datasets tailored to a range of NLP tasks and real-world use cases.
- The creation of **AveniBench**, a comprehensive collection of evaluation datasets specifically selected for financial services.
- The successful development of a series of increasingly powerful models (1B - 24B), culminating in the **FinLLM 24B M**, which outperforms strong public baselines on financial tasks, and the more efficient **FinLLM 14B M**, created via a pruning and distillation pipeline.
- The implementation and validation of a **synthetic data generation** pipeline as a powerful technique to overcome data scarcity and drastically improve performance in targeted capabilities like tabular reasoning.
- The demonstration of **real-world utility** in the Aveni Detect and Assist use cases, where FinLLM performed competitively against or surpassed larger, more expensive commercial models while simplifying workflows.
- The establishment of a comprehensive **safety and governance framework** and methods to ensure the responsible deployment of these powerful tools.

Future Outlook

From Model to Platform Capability

As FinLLM moves into its next phase, the focus shifts from model development to integration and scalability through the Aveni Agentic Platform - a modular and interactive interface to create and deploy agentic workflows.

FinLLM will no longer exist as a standalone research artefact; it will become a shared intelligence layer accessible through the platform's Task Registry and workflow engine. Each FinLLM model, whether for reasoning, classification, or information extraction, will be packaged as a callable Task within the platform, allowing it to power agentic workflows such as financial case analysis, compliance review, or customer insight generation.

Fine-Tuning module

A central pillar of the roadmap is the creation of the FinLLM Fine-Tuning Kit, deployed through the Aveni Platform. This toolkit will allow enterprises to fine-tune FinLLM safely and efficiently on their own proprietary data, within governed boundaries. The aim is to produce models that align with each institution's tone, risk policies, and compliance frameworks, while maintaining Aveni's ownership of FinLLM's sovereign IP and protecting the underlying model weights and data. This approach transforms FinLLM from a static model into a living capability that continuously learns from real-world usage.

Continuous Improvement and Expansion

In parallel, the FinLLM team will continue to expand the core model capabilities by introducing deeper reasoning, long-context understanding, improved safety alignment, and multilingual financial comprehension. The evaluation and guardrail layers of the Aveni Platform (including AveniBench and automated observability pipelines) will provide continuous feedback loops between model performance, fine-tuning data, and real-world task outcomes. Each iteration of FinLLM will therefore improve both the model itself and the workflows it powers.

Purpose and Vision

The purpose of FinLLM is to serve as the sovereign intelligence engine for financial services automation - a model that can be trusted, adapted, and proven. Within the Aveni Platform, it will underpin a new generation of agentic systems that combine reasoning, assurance, and observability by design. FinLLM's roadmap is therefore twofold:

1. Empower enterprises to fine-tune and deploy their own compliant financial models using the Aveni Agentic Platform toolkit.
2. Enable Aveni to build and orchestrate increasingly capable, task-specific models that drive automation safely across Assist, Detect, and future agentic workflows.

In short, FinLLM is evolving from a standalone model into an ecosystem capability, one that strengthens Aveni's platform, accelerates customer innovation, and redefines how regulated institutions can deploy, control, and continuously improve AI within their own environments.

The next chapter of financial AI is being written now.

Join us in shaping how responsible intelligence powers the industry.

hello@aveni.ai

aveni